Introduction to Computational Physics

Monte Carlo Methods - I

Mohsen Sadr

Paul Scherrer Institut, Villigen
E-mail: andreas.adelmann@psi.ch,
mohsen.sadr@psi.ch

https://moodle-app2.let.ethz.ch/course/view.php?id=23214

Content I

7.1 Introductory Comments I

Monte Carlo (MC) methods

- are a natural choice in problems with the curse of high-dimensionality.
- introduce noise.

Among the questions that we try to answer here:

- How can we reduce noise?
- For which dimensions MC is worth trying?

7.1.1 Reminder I

- **Sample Space** Ω is the space of all possible outcomes for a random event.
- σ-**algebra** $\mathcal{F}$ is a non-empty collection of subsets of Ω s.t.
  1. $\emptyset \in \mathcal{F}$
  2. $F \in \mathcal{F} \implies F^C \in \mathcal{F}$
  3. $F_1, F_2, ... \in \mathcal{F} \implies \bigcup_i F_i \in \mathcal{F}$
- The pair $(\Omega, \mathcal{F})$ is called a **measurable space**.
- **Probability measure** P on $(\Omega, \mathcal{F})$ is $P : \mathcal{F} \to [0, 1]$ s.t.
  1. $P(\emptyset) = 0, \ P(\Omega) = 1$
  2. if $F_1, F_2, ... \in \mathcal{F}$ are disjoint, $P(\bigcup_i F_i) = \sum_i P(F_i)$.
- The triple $(\Omega, \mathcal{F}, P)$ is called a **probability space**.
- A probability space is **complete** if its σ-algebra $\mathcal{F}_{\text{tot}}$ contains all possible subsets of Ω.
- **Random Variable** X is a measurable function $X : \Omega \to \mathcal{R}$.
- X induces a probability measure P_X and σ-algebra $\mathcal{F}_X \subset \mathcal{F}_{\text{tot}}$.

7.1.1 Reminder II

- **Law of Large Numbers (LLN):** Given a sequence of independent and identically distributed (iid) random variables $X_1, X_2, ..., X_N$ with $\mathbb{E}[X_i] = \mu$ for $i = 1, ..., N$,

$$\frac{1}{N}\sum_i X_i \to \mu \quad \text{as} \quad N \to \infty \ .$$

- **Central Limit Theorem (CLT):** Given a sequence of independent and identically distributed (iid) random variables $X_1, X_2, ..., X_N$ with $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$,

$$Z = \frac{S - N\mu}{\sigma\sqrt{N}} \sim \mathcal{N}(0,1) \quad \text{as} \quad N \to \infty,$$

where $S = \sum_{i=1}^{N} X_i$.

7.1.1 Reminder III

- Given finite and independent samples of a probability density function f, i.e. $\{X_{i=1}^{N}\} \sim f$, the expectation of an integrable function in this measure, i.e.

$$\mathbb{E}[\phi(X)] = \int \phi(x) f(x) dx,$$

can be estimated using the law of large numbers

$$\mathbb{E}^{\Delta}[\phi(X)] = \frac{\sum_i \phi(X_i)}{N} .$$

- Variance of this estimate (either using LLN or CLT) is

$$\mathrm{Var}(\mathbb{E}^{\Delta}[\phi(X)]) = \frac{\mathrm{Var}(\phi(X))}{N} ,$$

which is dimension independent.

7.1.2 Signature of MC Methods I

- Instead of discretizing the probability space deterministically, MC generates samples of the target distribution function.
- This allows exploring only relevant parts of phase space.
- Unbiased MC estimate converges towards the correct solution by increasing the number of samples.

Common Steps of Acceptance/Rejection Monte Carlo Methods

1. Randomly generate a new configuration.
2. Accept or reject the new configuration.
3. Sample the physical quantity of interest.
4. Repeat the process (go to 1) till enough samples are generated.

7.2 Statistical noise in standard Monte Carlo method I

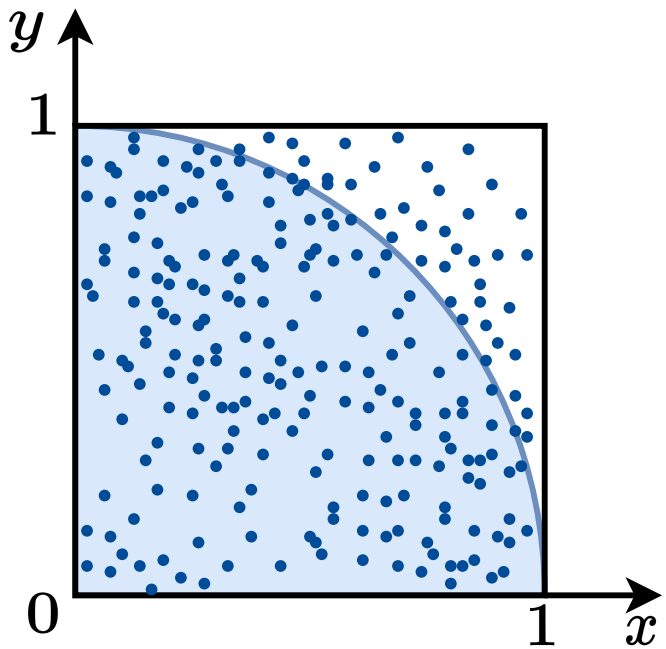
- The Monte Carlo methods rely on sampling and averaging (LLN) to estimate the solution.
- Hence, standard MC introduces statistical noise ϵ in prediction which reduces with the number of samples N via

$$\epsilon \propto \frac{1}{\sqrt{N}}$$

- The statistical noise is dimension independent.
- The slow convergence rate motivates variance reduction methods.

7.2.1 Computation of π I

Consider the unit square $S = \{(x, y); x \in [0, 1], y \in [0, 1]\}$ and a quarter circle $P = \{(x, y); x^2 + y^2 \leq 1, x \in [0, 1], y \in [0, 1]\}$. Given analytical solution of their areas $A_S = 1$ and $A_P = \pi/4$. we can have $\pi = 4A_P/A_s$.



7.2.1 Computation of π II

- Generate N random points $\{(x_i, y_i)\}_{i=1}^{N}$ uniformly distributed in the unit square S.
- For the ith point, if $x_i^2 + y_i^2 \leq 1 \Rightarrow N_c = N_c + 1$.

Then, the number of points N_c lying within the quarter circle is compared to the total number N of points and the fraction will give us an approximate value of π

$$\mathbb{E}^{\Delta}[\pi] = 4\frac{N_c(N)}{N}.$$

It can be observed

$$\epsilon = |\mathbb{E}^{\Delta}[\pi] - \pi| \propto \frac{1}{\sqrt{N}}.$$

7.3 Computation of Integrals I

- Consider $\int_a^b g(x)dx$.
- Integral is approximated numerically using N points X_i on the x-axis with their corresponding values $g(X_i)$.
- Summing and averaging over these sampled results and multiplying the resulting expression with the length of the interval

$$\int_a^b g(x)dx \approx (b-a)\left[\frac{1}{N}\sum_{i=1}^{N} g(X_i)\ .\right]$$

- What should be the distribution of $\{X_i\}_{i=1}^N$ on the x-axis?
- For $X \sim \mathrm{Uniform}(a,b)$, the integration method is called "**simple sampling**" which works well if $g(x)$ does not have large variations.

But what if

- $g(x)$ has large variations?

7.3.1 Integration Errors I

Error in Conventional Methods - e.g. Trapezoidal Rule

Consider the Taylor Series expansion integrated from x_0 to $x_0 + \Delta x$:

$$\begin{aligned}
\int_{x_0}^{x_0+\Delta x} f(x)dx &= f(x_0)\Delta x + \frac{1}{2}f'(x_0)\Delta x^2 + \frac{1}{6}f''(x_0)\Delta x^3 + ... \\
&= \left[\frac{1}{2}f(x_0) + \frac{1}{2}(f(x_0) + f'(x_0)\Delta x + \frac{1}{2}f''(x_0)\Delta x^2 + ...) + ...\right]\Delta x \\
&= \frac{1}{2}(f(x_0) + f(x_0 + \Delta x))\Delta x + \mathcal{O}(\Delta x^3) \\
&\rightarrow \text{ error in each interval } \propto (\Delta x)^3
\end{aligned}$$

7.3.1 Integration Errors II

Error in Conventional Methods - e.g. Trapezoidal Rule

We now subdivide our interval $[x_0, x_1]$ into N times $\Delta x = \frac{x_1 - x_0}{N}$.

$$\begin{aligned}
\int_{x_0}^{x_1} f(x)dx &= \sum_{j=0}^{N-1} \int_{x_j}^{x_j+\Delta x} f(x)dx \\
&= \sum_{j=0}^{N-1} \left[\frac{\Delta x}{2} \left(f(x_j) + f(x_j + \Delta x) \right) + \mathcal{O}(\Delta x^3) \right] \\
&= \sum_{j=0}^{N-1} \left[\frac{\Delta x}{2} \left(f(x_j) + f(x_j+) \right) \right] + \mathcal{O}(\Delta x^2) \\
&\rightarrow \text{total error is } \mathcal{O}(\Delta x^2) \propto \mathcal{O}(N^{-2}) \ .
\end{aligned}$$

The generalization to d-dimensions is described in the script, the result is:

7.3.1 Integration Errors III

Error in Conventional Methods - e.g. Trapezoidal Rule

Dimension dependent error

The error of 2nd order conventional deterministic integration methods $\propto (\Delta x)^2 \propto N^{-\frac{2}{d}}$

7.3.1 Integration Errors I

Monte Carlo Error

Consider a simplified case of a one dimensional function of one variable, $g : [a, b] \to \mathbb{R}$. If we pick N equidistant points in the interval $[a, b]$ we have a distance of $h = \frac{b-a}{N}$ between each of these points.
The estimate for the integral is

$$I = \int_a^b g(x)dx \approx \frac{b-a}{N} \sum_{i=1}^{N} g(x_i) = (b-a)\langle g \rangle \equiv Q$$

where $\langle g \rangle$ stands for the sample mean of the integrand.
The unbiased variance reads

$$\mathrm{Var}(g) \equiv \sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (g(x_i) - \langle g \rangle)^2$$

7.3.1 Integration Errors II

Monte Carlo Error

Using the Central Limit Theorem (CLT), the variance of the estimate of the integral can be computed as

$$\mathrm{Var}(Q) = (b-a)^2 \frac{\mathrm{Var}(g)}{N} = (b-a)^2 \frac{\sigma^2}{N}$$

which for large N decreases like $\frac{1}{N}$. Thus, the statistical noise is

$$\epsilon \approx \sqrt{\mathrm{Var}(Q)} = (b-a)\frac{\sigma}{\sqrt{N}}$$

We can now generalize this to multidimensional integrals:

$$I = \int_{a_1}^{b_1} dx_1 \int_{a_2}^{b_2} dx_2 ... \int_{a_n}^{b_n} dx_n f(x_1, ..., x_n)$$

7.3.1 Integration Errors III

Monte Carlo Error

The previous interval $[a, b]$ becomes a hypercube V as integration volume, with

$$V = \{x : a_1 \leq x_1 \leq b_1, ...a_n \leq x_n \leq b_n\}$$

Instead of the interval $[a, b]$ we now use the hypercube V

$$\mathrm{Var}(Q) = V^2 \frac{\mathrm{Var}(g)}{N} = V^2 \frac{\sigma^2}{N}$$

and the statistical noise becomes

$$\epsilon \approx \sqrt{\mathrm{Var}(Q)} = V \frac{\sigma}{\sqrt{N}}$$

Dimension independent error

The error in simple Monte Carlo methods scales as $\frac{1}{\sqrt{N}}$ independent of the dimension d.

7.3.1 Integration Errors I

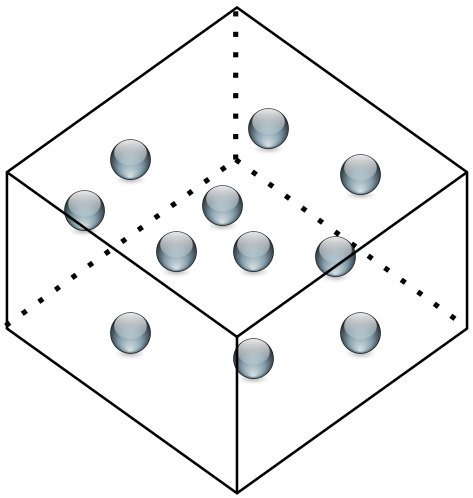
Comparison of the Errors

We have seen that in conventional methods, the error of 2nd order integral scheme goes with $N^{-\frac{2}{d}}$ and thus depends on the dimension, while the error in Monte Carlo methods is independent of the dimension. There is a crucial point at which Monte Carlo methods become more efficient,

$$N^{-\frac{2}{d}} \overset{crit}{=} \frac{1}{\sqrt{N}} \qquad \Rightarrow d_{crit} = 4$$

We can thus conclude that for $d > 4$, Monte Carlo becomes more efficient than 2nd order integral scheme and is therefore used in areas where such higher dimensional integrals with $d > 4$ are commonplace.

7.3.2 Higher Dimensional Integrals I



Let us now look at an example of higher dimensional integration: Consider N_p hard spheres of radius R in a 3D box of volume V. Our points are characterized by their position vector $\vec{x}_i = (x_i, y_i, z_i), 1 \leq i \leq N_p$. We define the distance between two such points as

$$r_{ij} := \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

where we have to take into consideration that the spheres are hard, i.e. cannot overlap. Thus the minimal distance between two neighboring spheres is the distance when the spheres are in contact, which is equal to $2R$. This translates to the following condition

$$r_{ij} \overset{!}{>} 2R \ .$$

7.3.2 Higher Dimensional Integrals II

This condition leads to the joint conditional probability density function for N_p particle system

$$f = f(\vec{x}_1, ..., \vec{x}_{N_p};\ ||\vec{x}_i - \vec{x}_j||_2 > 2R \quad \forall i, j = 1, ...N_p)\ .$$

Let us say we are interested in the average distance between the centers of spheres in the box. Then we need estimate the following integral

$$\mathbb{E}[r] = \int \left(\frac{1}{N_p(N_p-1)} \sum_{i \neq j} r_{ij} \right)\ f\ d^3\vec{x}_1 ... d^3\vec{x}_{N_p}.$$

Given N realizations (configurations/samples) of N_p particles system in the box, the average distance $\mathbb{E}[r]$ can be estimate via

$$\mathbb{E}^{\Delta}[r] = \frac{1}{N} \sum_{k=1}^{N} \left(\frac{2}{N_p\,(N_p-1)} \sum_{i<j} r_{ij} \right)\ .$$

7.3.2 Higher Dimensional Integrals III

The Monte Carlo solution to this formula is relatively simple:

- Choose a particle position (i.e. the center of the new sphere).
- Make sure that the new sphere does not overlap with any pre-existing spheres (see condition on r_{ij}). If it does overlap, reject the position and try again.
- Once all the spheres have been placed, calculate the distances r_{ij}.

We then use these distances r_{ij} to compute the average.

7.4 Utilizing Nonuniform Random Numbers I

Assumption: random numbers are uniformly distributed in a d-dim box.

This integration method can become expensive if f fills only a small fraction of the box.

- Find some other distribution which better encloses the volume we are integrating over, and generate random numbers only in this
- If the enclosing function is such that we know how to generate random numbers in this distribution analytically, the savings in time can be significant.

7.4.1 Importance Sampling I

Idea

Importance sampling attempts to reduce the statistical error $\epsilon \propto \frac{\sigma}{\sqrt{N}}$ by reducing the variance σ^2.

We note the following problem in the standard MC method:

- Consider integral of a function with large peaks.
- Most of uniformly distributed samples will not be in the peak area.

The idea behind importance sampling is to

- Transform the integrand into a flatter function.

7.4.1 Importance Sampling II

Assume we have a probability density function $g(x)$ with the property

$$\frac{f(x)}{g(x)} < \infty, \quad g(x) > 0, \forall x.$$

Hence the integral can be rewritten as

$$I = \int f(x)dx = \int \frac{f(x)}{g(x)} g(x)dx.$$

Now, $f(x)/g(x)$ acts as the function we want to compute its integral with the probability measure of $g(x)$.

Given N i.i.d. samples of $g(x)$, i.e. $\{X_i\}_{i=1}^{N} \sim g(x)$, this integral can be estimated via

$$I \approx \frac{1}{N} \sum_{i=1}^{N} f(X_i)/g(X_i) \ .$$

7.4.2 Control Variate I

- Similar to importance sampling for variance reduction.
- Integrate a flatter function.

Instead of division, use subtraction:

$$I = \int f(x)dx = \int (f(x) - g(x))dx + \int g(x)dx \tag{1}$$

Idea

- Find a $g(x)$ that $\mathrm{Var}(\mathbb{E}^{\Delta}[f(X) - g(X)]) < \mathrm{Var}(\mathbb{E}^{\Delta}[f(X)])$.
- $\int g(x)dx$ is known analytically .

The integral $\int (f(x) - g(x))dx$ is evaluated with normal MC sampling.
The following advantages compared to importance sampling are obvious:

- $g(x)$ can be zero at any point x.
- We do not need to generate samples of $g(x)$.

7.4.2 Control Variate II

Now the term $\mathrm{Var}(\mathbb{E}^{\Delta}[f(X) - g(X)])$ is

$$\begin{aligned}\mathrm{Var}(\mathbb{E}^{\Delta}[f(X) - g(X)]) =& \mathrm{Var}(\mathbb{E}^{\Delta}[f(X)]) + \mathrm{Var}(\mathbb{E}^{\Delta}[g(X)]) \\ & - 2\mathrm{Cov}(\mathbb{E}^{\Delta}[f(X)], \mathbb{E}^{\Delta}[g(X)])\end{aligned} \tag{2}$$

where the last term is the covariance between estimates of f and g. This method is only practically useful when

$$2\ \mathrm{Cov}(\mathbb{E}^{\Delta}[f(X)], \mathbb{E}^{\Delta}[g(X)]) \approx \mathrm{Var}(\mathbb{E}^{\Delta}[f(X)]). \tag{3}$$

which translates into estimates of f and g being positively correlated.

7.5 Quasi-Monte Carlo I

$$I = \frac{1}{N} \sum_{i=1}^{N} f(x_i) \tag{4}$$

with $x_1, \ldots, x_N$ deterministic (and cleverly chosen). How to choose $x_1, \ldots, x_N$?

- lattice rules (many of them exist)
- low discrepancy points (see chapter on random numbers)

The error of this approximation scales as

$$\mathcal{O}\left(\frac{(\log N)^d}{N}\right).$$

7.6 Markov Chain Monte Carlo I

Reminder: A **Stochastic Process** is a parameterized collection of random variables $\{X_t\}_{t\in[0,T]}$ defined on $(\Omega, \mathcal{F}, P)$.

Idea

A Markov Chain is a stochastic process that is used to predict/estimate/guess the outcome of an event given only the previous state and its action.
If a sequence of events exhibits the this property, then the process is called "Markovian" in nature.

Applications of Markov Chain Monte Carlo:

- Sampling a target density.
- Simulating/modelling physical processes.

7.6 Markov Chain Monte Carlo II

A Markov chain X is a sequence $X_1, X_2, \ldots$ of stochastic variables, which for all $n > 0$ and all events $A_1, A_2, \ldots, A_n$ satisfies the following conditional independence property:

$$P(X_n \in A_n | X_{n-1} \in A_{n-1}, \ldots, X_0 \in A_0) = P(X_n \in A_n | X_{n-1} \in A_{n-1})$$

Example: We consider a Markov chain on the discrete state sample space $\Omega = \{\text{Ground State}, \text{Exited State}\} = \{G, E\}$.
A Markov chain X on $\Omega = \{G, E\}$ is determined by the initial distribution given by

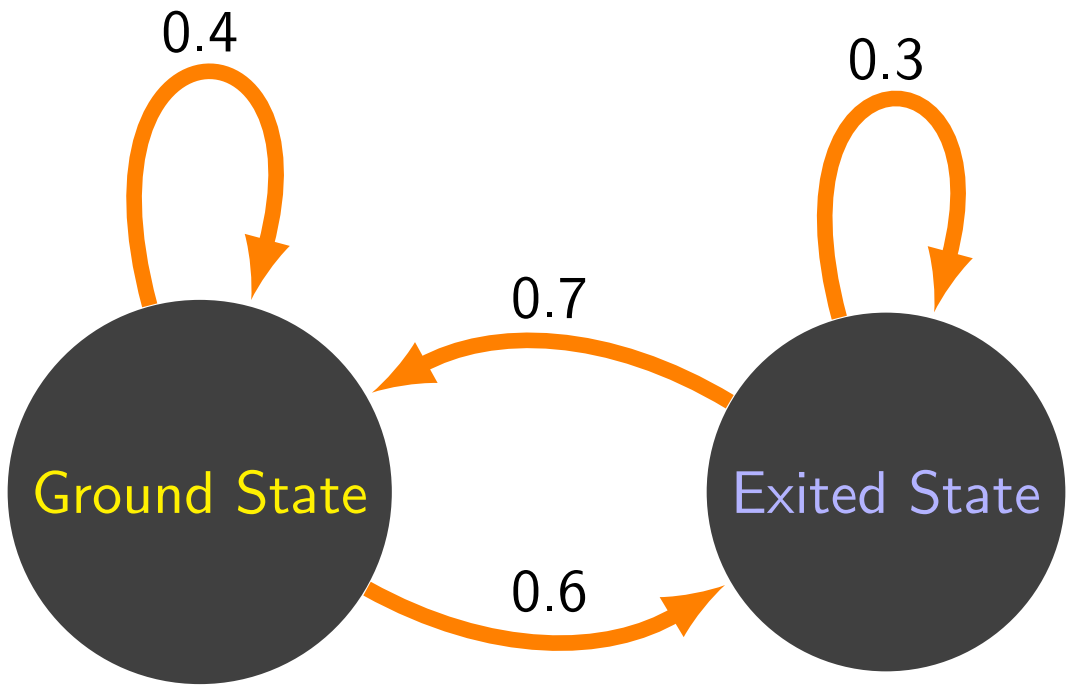
$$p_0 = P(X_n = G) \text{ and } p_1 = P(X_n = E)$$

and the one-step transition probabilities given by

$$p_{00} = P(X_{n+1} = G | X_n = G), \quad p_{10} = P(X_{n+1} = G | X_n = E)$$

$$p_{01} = 1 - p_{00}, \quad p_{11} = 1 - p_{10}$$

7.6 Markov Chain Monte Carlo III



The one-step transition probabilities can be written as a matrix

$$P = \begin{pmatrix} p_{00} & p_{10} \\ p_{01} & p_{11} \end{pmatrix} = \begin{pmatrix} 0.4 & 0.6 \\ 0.7 & 0.3 \end{pmatrix}$$

7.6.6 Canonical Monte Carlo I

What is an Ensemble?

- An ensemble is the collection of identically distributed and independent systems.
- One usually considers the phase space (which for N_p particles is $6N_p$ dimensional).
- Selected points are then regarded as a collection of representative points in phase space.
- The normalizing factor of the measure is called the partition function of the ensemble
- An ensemble is said to be stationary if the associated measure is time-independent.

7.6.6 Canonical Monte Carlo II

What is an Ensemble?

The most important ensembles are the following:

- Microcanonical ensemble: (E, V, N) constant,
- Canonical ensemble: (T, V, N) constant,
- Grand-canonical ensemble: (T, V, μ) constant.

Space average (ensemble average) of an observable f defined on phase space Λ with the probability measure $d\mu$ is

$$\langle f \rangle = \int_{\Lambda} f d\mu \ .$$

Time average is

$$\bar{f}_t = \lim_{T \to \infty} \frac{1}{T} \int_0^T f(x(t)) dt \ .$$

In a **Ergodic process**

$$\langle f \rangle = \bar{f}_t \ .$$

7.6 Back to Markov Chain Monte Carlo I

Let the energy of configuration X be given by $E(X)$. Then, the distribution of particle velocity at equilibrium is given by the Boltzmann distribution

$$p_{eq}(X) = \frac{1}{Z_T} e^{-\frac{E(X)}{k_B T}} \text{ with } \sum_X p_{eq}(X) = 1$$

and Z_T is the partition function (the normalizing factor of the measure) :

$$Z_T = \sum_X e^{-\frac{E(X)}{k_B T}} \ .$$

Let us now consider an ensemble average which for a discrete X reads

$$\langle Q \rangle = \sum_X Q(X) p_{eq}(X) \ .$$

7.6 Back to Markov Chain Monte Carlo II

Target: Calculate the ensemble average for a given property, e.g. energy.

Challenge: We need many samples that satisfy the constraints.

Solution: Markov Chain Monte Carlo provides an efficient method for sampling appropriate configuration out of a large pool of possibility.

Setup:

1. Introduce a virtual time τ for the stochastic process.
2. Start in a given configuration X.
3. Propose a new configuration Y with a transition probability $T(X \to Y)$.
4. Normalize over all possible new configurations $\sum_Y T(X \to Y) = 1$.
5. go to 3 until convergence.

7.6 Back to Markov Chain Monte Carlo III

The probability of state being at X given the Markov Chain Monte Carlo process follows the **Master equation**

$$\frac{dp(X,\tau)}{d\tau} = \sum_{Y \neq X} p(Y)T(Y \to X) - \sum_{Y \neq X} p(X)T(X \to Y) \ .$$

Theorem

A Markov chain that satisfies

1. Ergodicity: $\forall X, Y : T(X \to Y) > 0$
2. Normalization: $\sum_Y T(X \to Y) = 1$
3. Homogeneity: $\sum_Y p(Y)T(Y \to X) = p(X)$

converges towards a unique stationary distribution p_{st}.

7.6 Back to Markov Chain Monte Carlo IV

To see this, we set the stationary distribution p_{st} equal to the equilibrium distribution of the considered physical system p_{eq}

$$\frac{dp(X,\tau)}{d\tau} = 0 \Leftrightarrow p_{st} \stackrel{!}{=} p_{eq}.$$

Detailed Balance

This leads to the concept of *Detailed Balance* and tells us that the steady state of the Markov process in the thermal equilibrium. We have achieved that by using the Boltzmann distribution for $p(X)$.

Details can be found in the script.

7.6.2 M(RT)^2 Algorithm I

- M(RT)^2 is an abbreviation of the last names of the authors of the original paper.
- RT is squared because except Metropolis, the other four authors (Rosenbluth, Teller) of the paper formed two married couples.
- They introduced a Markov Chain Monte Carlo method for sampling equilibrium distribution function.
- Later W. Hastings extended it to general distributions, commonly known as Metropolic-Hastings algorithm.

7.6.2 M(RT)^2 Algorithm II

In the Metropolis algorithm, the acceptance probability A is defined as

$$A(X \to Y) = \min\left(1, \frac{p_{eq}(Y)}{p_{eq}(X)}\right)$$

We can now insert the Boltzmann distribution

$$p_{eq}(X) = \frac{1}{Z_T} e^{-\frac{E(X)}{k_B T}}$$

to find

$$A(X \to Y) = \min\left(1, e^{-\frac{E(Y)-E(X)}{k_B T}}\right) = \min\left(1, e^{-\frac{\Delta E}{k_B T}}\right)$$

Suppose we go to a configuration of lower energy, ΔE will be negative and we would end up with an exponential expression like $\exp(\frac{|\Delta E|}{k_B T})$, at which point the minimum kicks in and sets the expression to one. In other words,

7.6.2 M(RT)^2 Algorithm III

the acceptance will be equal to one ("always accept") for transitions to configurations of lower energy.
If we go to a configuration of higher energy, ΔE will be positive and we'll have $\exp(-\frac{|\Delta E|}{k_B T}) < 1$ so the acceptance will increase with the temperature.

Remark

Note that a thermal equilibrium is enforced by detailed balance and we impose that the steady state must be a Boltzmann distribution.

7.6.3 The Ising Model I

- The Ising model is a model that originally aimed at explaining ferromagnetism.
- Today used in many other areas such as opinion models, binary mixtures, lattice gas ....
- A highly simplified approach to e.g. the difficulties of magnetism (e.g. commutation relations of spins).

Setup:

- Consider a discrete collection of N binary variables called spins.
- They take on the values ± 1 (representing the up and down spin configurations).
- The spins σ_i interact pairwise.
- The energy has one value for aligned spins ($\sigma_i = \sigma_j$) and another for anti-aligned spins ($\sigma_i \neq \sigma_j$).

7.6.3 The Ising Model II

The Hamiltonian is given by

$$\mathcal{H} = -\sum_{ij} J_{ij}\sigma_i\sigma_j - \sum_i H_i\sigma_i \tag{5}$$

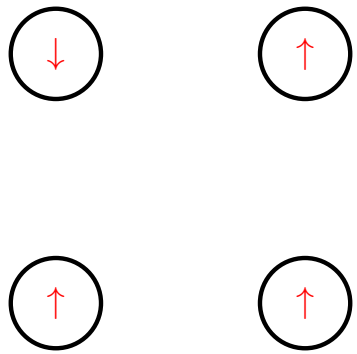
where H_i is a (usually homogeneous) external field and J_{ij} are the (translationally invariant) coupling constants. As the coupling constants are translationally invariant, we may drop the indices and simply write $J = J_{ij} \ \forall i, j$. The coupling constant J is half the difference in energy between the two possibilities (alignment and anti-alignment).

The simplest example is the antiferromagnetic one-dimensional Ising model, which has the energy function

$$E = \sum_i \sigma_i\sigma_{i+1} \tag{6}$$

7.6.3 The Ising Model III

which can be generalized to the two-dimensional case; we can also add an external field as in equation (5).
Consider for illustration an ensemble of 4 particles with spins i.e., $|S| = 2^4 = 16$ states $X_1, \ldots, X_{|S|}$. An example of a particular state is $X_j = (\downarrow, \uparrow, \uparrow, \uparrow)$ is depicted below.



7.6.3 The Ising Model IV

A Markov chain is generated by stepping from X_i to X_{i+1}, with probability $p_{i,i+1}$ or short $p_{i,j}$ with $j = i + 1$. The probability $p_{i,j}$ is called the transition probability. With $\boldsymbol{P}$ the transition matrix comprised of $p_{i,i+1}$

$$p_{i,j} = X_i^T \boldsymbol{P} X_j, \quad \dim \boldsymbol{P} = |S| \times |S|. \tag{7}$$

The following normalizing condition must be fulfilled:

$$p_i = \sum_{j=1}^{|S|} p_{i,j} = 1 \tag{8}$$

were p_i means the probability to any state i.
A Markov process creates a sequence of states $X^{(\tau)}$ that are labeled by the *virtual/algorithmic time* (τ).

7.6.3 The Ising Model V

The M(RT)^2 Algorithm & the Ising Model

1. Randomly choose a configuration X_i
2. Compute $\Delta E = E(Y) - E(X) = 2J\sigma_i\sigma_j$
3. Spinflip if $\Delta E < 0$, otherwise accept with prob. $\exp\left(-\frac{\Delta E}{k_B T}\right)$

Introduction to Computational Physics

Monte Carlo Methods - A glimpse on Multilevel Monte Carlo (MLMC)

Mohsen Sadr

Paul Scherrer Institut, Villigen
E-mail: andreas.adelmann@psi.ch,
mohsen.sadr@psi.ch

https://moodle-app2.let.ethz.ch/course/view.php?id=23214

Reminder I

- A **Stochastic Process** is a parameterized collection of random variables $\{X_t\}_{t\in[0,T]}$ defined on $(\Omega, \mathcal{F}, P)$.
- At a given t, we have a random variable $\omega \to X_t(\omega),\ \omega \in \Omega$, and for a fixed $\omega \in \Omega$ we have a **path** $t \to X_t(\omega), t \in [t_0, T]$.
- A **Wiener process** W_t for $t \geq 0$ is a process s.t.:
  1. $W_0 = 0$.
  2. W_t has independent increments.
  3. $W_t - W_s \sim \mathcal{N}(0, |t-s|)$.
  4. W_t is continuous in time.
- A **Stochastic Differential Equation (SDE)** is a differential equation that contains at least one stochastic process.
- A **drift-diffusion** SDE can be constructed simply by $dX_t = a(X_t)dt + b(X_t)dW_t$.
- W_t is not differentiable. In this notation $dW_t := W_{t+\Delta t} - W_t$.

Reminder II

- Drift-diffusion process can be solved numerically, e.g. using **Euler Maruyama scheme**

$$X_{t+\Delta t} = X_t + a(X_t)\Delta t + b(X_t)\xi, \text{ where } \xi \sim \mathcal{N}(0, \Delta t).$$

- **Law of Large Numbers (LLN):** Given a sequence of independent and identically distributed (iid) random variables $X_1, X_2, ..., X_n$ with $\mathbb{E}[X_i] = \mu$ for $i = 1, ..., n$,

$$\frac{1}{n}\sum_i X_i \to \mu \quad \text{as} \quad n \to \infty \ .$$

Reminder III

- **Central Limit Theorem (CLT):** Given a sequence of independent and identically distributed (iid) random variables $X_1, X_2, ..., X_n$ with $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$,

$$Z = \frac{S - n\mu}{\sigma\sqrt{n}} \sim \mathcal{N}(0,1) \quad \text{as} \quad n \to \infty,$$

  where $S = \sum_{i=1}^{n} X_i$.
- Given finite and independent samples of a probability density function f, i.e. $\{X_{i=1}^{n}\} \sim f$, the expectation of an integrable function in this measure, i.e.

$$\mathbb{E}[\phi(X)] = \int \phi(x) f(x) dx,$$

  can be estimated using the law of large numbers

$$\mathbb{E}^{\Delta}[\phi(X)] = \frac{\sum_i \phi(X_i)}{n} .$$

Reminder IV

- Variance of this estimate (either using LLN or CLT) is

$$\mathrm{Var}(\mathbb{E}^{\Delta}[\phi(X)]) = \frac{\mathrm{Var}(\phi(X))}{n} \ ,$$

  which is dimension-independent.

7.7 Multilevel Monte Carlo (MLMC) I

Source: Giles, Michael B. "Multilevel Monte Carlo methods." Acta Numerica 24 (2015): 259-328.

Consider the problem of estimating

$$\mathbb{E}[P] = \int p \; f \; dx$$

where $P : \Omega \to \mathbb{R}$ is a random variable of interest and f is the associated probability density.

Given N i.i.d samples $\{\omega\}_{i=1}^{N} \in \Omega$, LLN provides us with the unbiased estimate

$$\mathbb{E}^{\Delta}[P] = \frac{\sum_{i=1}^{N} P(\omega_i)}{N}$$

with the variance

$$\mathrm{Var}(E^{\Delta}[P]) = \frac{\mathrm{Var}(P)}{N} \; .$$

7.7 Multilevel Monte Carlo (MLMC) II

- Assume we can generate samples of P at different levels of cost versus accuracy.
- Samples obtained on the most accurate (finest) predictor are the most expensive.
- And samples from the least accurate (coarsest) predictor are the least expensive.
- **Question**: Can we combine them to obtain an estimate that is as accurate as the finest predictor and as cheap as the coarsest predictor?

7.7 Multilevel Monte Carlo (MLMC) III

- Let us denote samples of P at levels $l = l_0, ..., L$ with P_l. Here, accuracy and cost increases with l.
- Consider the telescopic sum

$$\mathbb{E}^{\Delta}_{\mathrm{MLMC}}[P] = \mathbb{E}^{\Delta}[P_{l_0}] + \sum_{l=l_0+1}^{L} \mathbb{E}^{\Delta}[P_l - P_{l-1}] \, . \tag{9}$$

- Clearly, as we increase number of samples

$$\mathbb{E}^{\Delta}_{\mathrm{MLMC}}[P] \to \mathbb{E}[P_L] \, . \tag{10}$$

- If we can find correlated sampled between every two levels such that

$$\mathrm{Var}(\mathbb{E}^{\Delta}[P_l - P_{l-1}]) \leq \mathrm{Var}(\mathbb{E}^{\Delta}[P_{l_0}]), \tag{11}$$

  then

$$\mathrm{Var}(\mathbb{E}^{\Delta}_{\mathrm{MLMC}}[P]) \approx \mathrm{Var}(\mathbb{E}^{\Delta}[P_{l_0}]) \, . \tag{12}$$

7.7 Multilevel Monte Carlo (MLMC) IV

Let C_l, V_l denote cost and variance of one sample of P_{l_0} for $l = l_0$ and $P_l - P_{l-1}$ for $l > l_0$.

Given N_l number of samples per level, we have the unbiased estimator

$$\mathbb{E}^{\Delta}_{\mathrm{MLMC}} = \frac{\sum_{i=1}^{N_0} P_0(w_i)}{N_0} + \sum_{l=l_0+1}^{L} \left(\frac{\sum_{i=1}^{N_l} P_l(\omega_i) - P_{l-1}(\omega_i)}{N_l} \right) \tag{13}$$

We note:

- Total Cost is $\sum_{l=l_0}^{L} N_l C_l$.
- Total Variance is $\sum_{l=l_0}^{L} V_l / N_l$.

Question: How many levels and samples per level are needed to get variance and accuracy of ϵ^2?

7.7 Multilevel Monte Carlo (MLMC) V

Consider an optimization problem with the cost function $\mathcal{C}$ that aims to minimize the cost of estimating the target expectation $\sum_{l=l_0}^{L}(N_l C_l)$ with constraint on the outcome variance to be ϵ^2, i.e.

$$\mathcal{C} = \sum_{l=l_0}^{L}(N_l C_l) + \mu^2 \left[\sum_{l=l_0}^{L}\left(\frac{V_l}{N_l}\right) - \epsilon^2\right], \tag{14}$$

where μ is the Lagrange multiplier. Note that in the original paper, the desired variance ϵ^2 is not incorporated in the constraint, which I think is a typo.

By taking derivative of $\mathcal{C}$ w.r.t. N_l, i.e. $\partial\mathcal{C}/\partial N_l$, we find the extremum at

$$\frac{\partial\mathcal{C}}{\partial N_l} = 0 \implies \sum_{l=l_0}^{L}\left(C_l - \mu^2\frac{V_l}{N_l^2}\right) = 0 \implies N_l^* = \mu\sqrt{\frac{V_l}{C_l}}\,. \tag{15}$$

7.7 Multilevel Monte Carlo (MLMC) VI

Next, we enforce the constraint on the extremum, i.e. letting $\sum V_l/N_l^* = \epsilon^2$ at the extremum, Lagrange multiplier can be found as a function of ϵ, i.e.

$$\epsilon^2 = \sum_{l=l_0}^{L} \frac{V_l}{N_l^*} = \sum_{l=l_0}^{L} \frac{V_l}{\mu\sqrt{V_l/C_l}} \implies \mu = \epsilon^{-2} \sum_{l=l_0}^{L} \sqrt{V_l C_l}\ . \tag{16}$$

The optimal number of samples per level then becomes

$$N_l^* = \mu\sqrt{\frac{V_l}{C_l}} = \epsilon^{-2} \left(\sum_{l=0}^{L} \sqrt{V_l C_l} \right) \sqrt{\frac{V_l}{C_l}}. \tag{17}$$

By substituting N_l^* and μ back in the total cost eq. 14, we obtain the total cost at the extremum as

$$\mathcal{C}^* = \epsilon^{-2} \left(\sum_{l=l_0}^{L} \sqrt{V_l C_l} \right)^2 . \tag{18}$$

7.7 Multilevel Monte Carlo (MLMC) VII

Remaining questions:

- How many levels are needed for a desired variance/accuracy?
- Do we need to know the variance in all levels a priori?
- What is the speedup compared to standard MC?

7.7 Multilevel Monte Carlo (MLMC) VIII

In order to test for weak convergence (error) w.r.t L, we need to check

$$|\mathbb{E}[P - P_L]| < \epsilon. \tag{19}$$

Assuming $\mathbb{E}^{\Delta}[P_l - P_{l-1}] \propto 2^{-\alpha l}$, the remaining error due to truncation of telescopic sum is

$$\mathbb{E}[P - P_L] = \sum_{L+1}^{\infty} \mathbb{E}[P_l - P_{l-1}] \tag{20}$$

$$\approx |\mathbb{E}^{\Delta}[P_L - P_{L-1}]|/(2^{\alpha} - 1) \tag{21}$$

which gives us a convergence condition for MLMC

$$|\mathbb{E}^{\Delta}[P_L - P_{L-1}]|/(2^{\alpha} - 1) < \epsilon. \tag{22}$$

Here, α is the convergence rate and is estimated either in advance or using linear regression during simulation.

7.7 Multilevel Monte Carlo (MLMC) IX

Algorithm 1 Pseudocode for MLMC

Initialize $L = 3$ and create N_0 samples on levels $l = 1, 2, 3$.
while not convergence **do**
    Draw extra samples for each level given N_l.
    Update estimate of V_l for $l = 1, ..., L$.
    Compute optimal N_l for $l = 1, ..., L$.
    if not converged and target N_l are already computed. **then**
        Set $L \leftarrow L + 1$ and update N_l for $l = 1, ..., L$.
    end if
end while

7.8 Example: Drift-Diffusion process I

- Consider an Ito process

$$dX = -X^p dt + \sqrt{2}\,\sigma dW \tag{23}$$

  where W is a standard Wiener process
- For simplicity, assume a constant diffusion coefficient $\sigma = 1$.
- Such process is used to model and simulate Brownian motion.
- Using Euler Maruyama scheme, one can discretize this SDE

$$X(t_0 + \Delta t) = X(t_0) - X^p(t_0)\Delta t + \sqrt{2\Delta t}\,\sigma\xi \tag{24}$$

  where $\xi \sim \mathcal{N}(0, 1)$.

Task: Estimate $\mathbb{E}[X]$ at time $t = T$ given initial value $X(t = 0)$.

7.8 Example: Drift-Diffusion process II

- We can construct solution at different levels of accuracy/cost using different time step sizes.
- At level l, we consider a Monte Carlo solution obtained using 2^l steps for the time interval $t \in [0, T]$.
- In order to sample $\mathbb{E}^{\Delta}[P_l - P_{l-1}]$, in each time step of coarse level $l-1$, two steps of fine level l needs to be done, i.e. on fine level

$$X_l \leftarrow X_l - X_l^p \Delta t_l + \sqrt{2\Delta t_l}\, \sigma \xi_1 \tag{25}$$

$$X_l \leftarrow X_l - X_l^p \Delta t_l + \sqrt{2\Delta t_l}\, \sigma \xi_2 \tag{26}$$

and coarse one

$$X_{l-1} \leftarrow X_{l-1} - X_{l-1}^p \Delta t_{l-1} + \sqrt{2\Delta t_{l-1}}\, \sigma \xi_3 \ . \tag{27}$$

Q: How can we maintain correlation between samples of l and $l-1$ levels?

7.8 Example: Drift-Diffusion process III

- We note that if ξ_1 and ξ_2 are i.i.d $\xi_1, \xi_2 \sim \mathcal{N}(0,1)$, then $\xi_1 + \xi_2 \sim \mathcal{N}(0,2)$.
- By setting $\xi_3 = (\xi_1 + \xi_2)/\sqrt{2}$, we have $\xi_3 \sim \mathcal{N}(0,1)$ correlated to ξ_1 and ξ_2.
- Reminder: Given independent random variables $X \sim f_X$ and $Y \sim f_Y$, then $X + Y \sim f_X * f_Y$.
- Reminder: Convolution of a normal distributions $\mathcal{N}(m_1, \sigma_1^2)$ with another normal $\mathcal{N}(m_2, \sigma_2^2)$, is a normal distribution, i.e. $\mathcal{N}(m_1, \sigma_1^2) * \mathcal{N}(m_2, \sigma_2^2) = \mathcal{N}(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$.

7.8 Example: Drift-Diffusion process IV

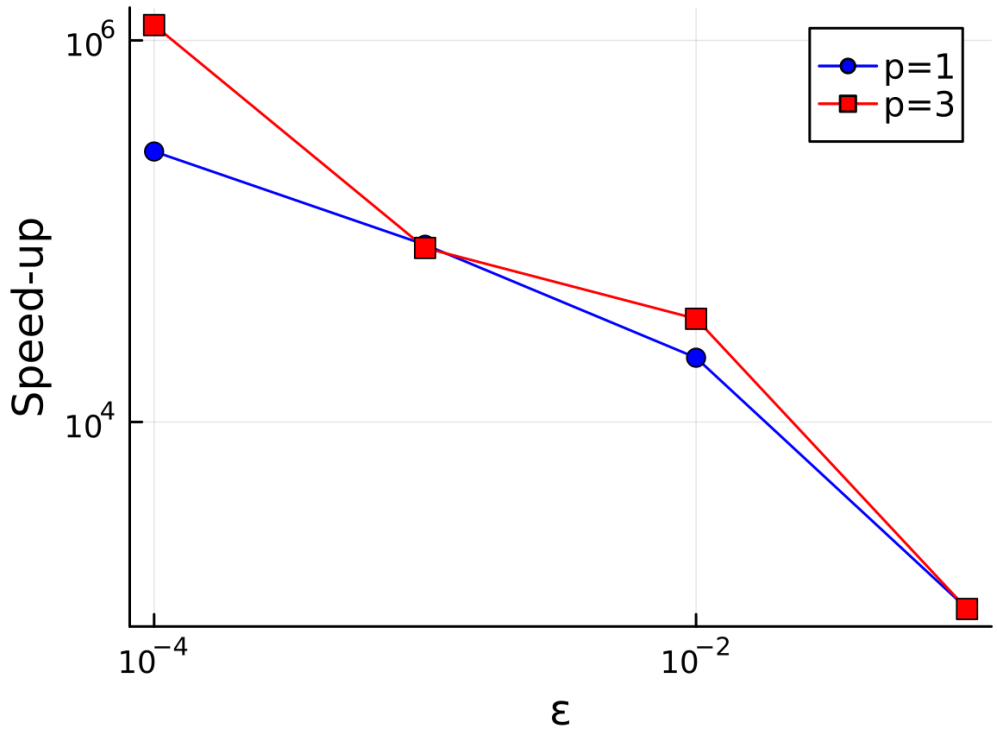


Figure: Speed-up of MLMC compared to standard MC for Langevin equation with $p = 1, 3$.